HMM Based Machine Translation

Team APK

Problem Statement

- English To Hindi Text Translation.
- HMM based Statistical Translation Model.

Input Data Set

- Tourism Text.
- 25 files of Parallel corpus.
- Each file has 1000 sentences.
- POS tags also given.

Preprocessing

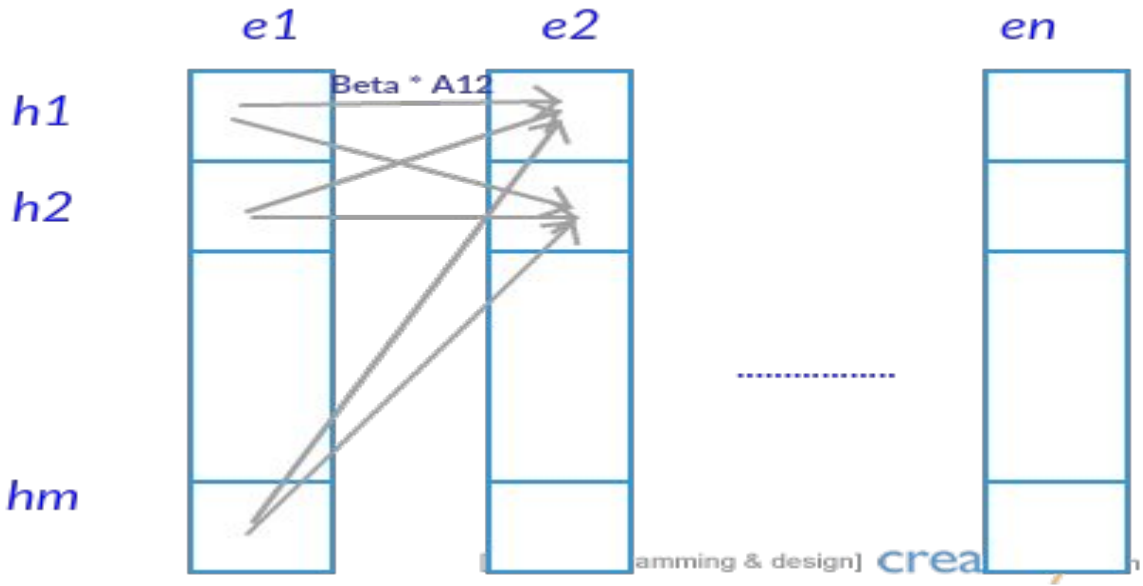
- Tokenize text
  - Space is considered as a delimiter for tokenisation.
  - Removed POS Tags
  - Hindi encoded as UTF-8

HMM Learning

- Emission Probabilities : Translation Probabilities.
  - $b_j(o_t)$: probability of observing o_t in state j.

- Transition Probabilities.
  - a_{ij} : probability of state transition from i to j.

HMM-Viterbi Algorithm



Method 1

- Co-occurrence Matrix for emission probabilities.
- Language Model for Transition Probabilities.

Method 2

- Aligner Used for emission Probabilities.
  - Nltk IBM2 aligner.

Method 3

- Transition probabilities calculated using both language model concepts and outputs generated from aligner.

Sample 1

94 Fair takes place on the banks of the Ganges in an area of about 11 kilometres .

94 मेला गंगा के किनारे लगभग 11 किलोमीटर के क्षेत्र में लगता है ।

घाट है लगता के शांति है लगता मुक्तेश्वर गढ़ में मेलों विशालतम हैं करते निवास है लगता

घाट लगभग किनारे किलोमीटर क्षेत्र 11 में गंगा के के तट के पर स्थान 11 11

घाट लगभग किनारे किलोमीटर के क्षेत्र 11 में गंगा के की तट के लगता 11 11 11

Sample 2

97 Close circuit TVs have also been installed at main places .

97 प्रमुख स्थानों पर क्लोज सर्किट टीवी कैमरे भी लगाए गए हैं ।

घाट भी कैमरे टीवी सर्किट क्लोज भी कैमरे टीवी सर्किट क्लोज

घाट स्थानों मुख्य दूरी टीवी सर्किट भी लगाई कैमरे प्रमुख प्रमुख

घाट स्थानों लगाए सर्किट टीवी सर्किट भी क्लोज कैमरे प्रमुख प्रमुख

Thank You.